Centre for Computational Finance and Economic Agents

WP034-08

Working Paper Series

Iacopo Giampaoli, Wing Lon Ng and Nick Constantinou

Analysis of Ultra-High-Frequency Financial Data Using Advanced Fourier Transforms

October 2008

University of Essex

CCFEA

www.ccfea.net

Analysis of Ultra-High-Frequency Financial Data Using Advanced Fourier Transforms [1]

Iacopo Giampaoli, Wing Lon Ng *, Nick Constantinou

Centre for Computational Finance and Economic Agents (CCFEA)
University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK

Abstract

This paper presents a novel application of advanced methods from Fourier analysis to the study of ultra-high-frequency financial data. The use of Lomb-Scargle Fourier transform, provides a robust framework to take into account the irregular spacing in time, minimizing the computational effort. Likewise, it avoids complex model specifications (e.g. ACD or intensity models) or resorting to traditional methods, such as (linear or cubic) interpolation and regular resampling, which not only cause artifacts in the data and loss of information, but also lead to the generation and use of spurious information.

Key words: Ultra-high frequency data, irregularly spaced data, Fourier Analysis.
JEL Code: C22, C63.

1 Introduction

Recently there has been a growing interest in the analysis of intraday data collected from financial e-markets. The increased availability of tick-by-tick "ultra-high frequency" (UHF) data fostered by the decreasing cost of data gathering, storing and manipulation, has certainly encouraged this trend leading to the analysis of financial data that was not readily available in the past,

* Corresponding author.
Email addresses: `igiamp@essex.ac.uk` (Iacopo Giampaoli), `wlng@essex.ac.uk` (Wing Lon Ng), `nconst@essex.ac.uk` (Nick Constantinou).

[1] Wing Lon Ng is grateful to the German Research Foundation (DFG) for financial support through research grant 531283. Nick Constantinou would like to thank HSBC for financial support.
The authors are very grateful to Richard Olsen for his helpful comments and suggestions.

requiring new techniques for efficient analysis. The availability of UHF data opens a window to provide new insights in the understanding of financial systems at the deepest market microstructure level.

UHF data (such as order book data) is observed in real-time and at non-periodic (irregular) intervals, typically resulting in a significant amount of information over a relatively short period. Although the literature has suggested several econometric models to deal with this specific data type, such as the ACD model (Engle and Russell, 1998) and its recent enhancement (for a comprehensive overview see Bauwens et al., 2004; Fernandes and Grammig, 2006; Bauwens and Hautsch, 2006a), count models (Grammig et al., 2004) or intensity models (see Hall and Hautsch, 2006; Bauwens and Hautsch, 2006b), none of these approaches seems apt to analyse the data both in real-time and within a multivariate framework. Duration models, as pointed out by Hall and Hautsch (2006), cannot consider more than a single process because of the asynchronization problem of multivariate point processes (otherwise truncation is required as shown in Engle and Lunde, 2003). Other approaches are easier to extend to multidimensional settings, but they either lose information (count models) or involve complex model specifications and significant computational time (intensity models). Furthermore, since point processes naturally focus on points, they sometimes overlook other variables and information recorded in the order book. In fact, though order book variates are sometimes included in the models as explanatory variables for duration or intensity, they are never modelled as independent self-contained processes.

The analysis of standard (periodic) econometric time series in the frequency domain, employing the Fast Fourier Transform (FFT) has a long history. In this Letter, it is demonstrated for the first time, that UHF data can also be analysed in the frequency domain via the so-called "Lomb-Scargle Fourier Transform" (LSFT) without the need for any data manipulation. The LSFT is the generalisation of the FFT to non-periodic processes and therefore provides a natural framework for the study of UHF data in the frequency domain. The LSFT has been successfully developed and applied in diverse areas of science, such as astronomy (Scargle, 1982, 1989), cardiology (*inter alia* Laguna et al., 1998; Castiglioni and Rienzo, 1996; Chang et al., 2001), palaeoclimatology (Schulz and Stattegger, 1997), meteorology (Muller and Macdonald, 2000), and biology (see Glynn et al., 2006, and Ruf, 1999). All these application of the LSFT share the property that the data are non-periodic, which is of course one of the defining properties of UHF data.

Indeed, a significant result of this simulation study is that resampling or interpolating UHF data to transform the set to a periodic data set for analysis, as has been carried out by other authors (see, for example, Dacorogna et al., 2001), introduces spurious data (or market microstructure noise) and loss of information. In order to clearly demonstrate this, the LSFT is applied to both periodic and non-periodic series generated by a benchmark AR(1) process, and a Monte Carlo simulation is performed to highlight its features and the advantages of this methodology over the traditional FFT coupled with regular resampling.

The outline of the Letter is as follows; Section 2 introduces the LSFT, Section 3 presents the simulation results and Section 4 provides the conclusions.

2 Spectral analysis of unevenly-spaced data

The general aim of spectral analysis is the decomposition of a time series into its periodic (frequency) constituents in order to detect and analyse the cyclical behaviour of economic processes by means of the estimation of the spectral density function (SDF) (see Priestley, 1981; a survey of spectral analysis of economic time series can be found in Granger and Engle, 1984 and Iacobucci, 2003). The SDF is the natural tool to analyse the frequency properties of time series, and is the analogue of the autocorrelation function in the time domain. It describes how the variation in a time series may be accounted for by components at different frequencies, and provides information on the contribution to the variance of the underlying process from different frequencies. Spectral analysis of evenly-spaced time series is typically carried out using a FFT algorithm (Priestley, 1981). UHF data are, however, unevenly-spaced, and hence the traditional FFT cannot be applied to the raw data. Attempts to transform the irregularly spaced raw data into regularly spaced data (e.g., Dacorogna et al., 2001), by regular resampling or using interpolation, prior to applying the FFT to calculate the SDF, is demonstrated in the following section to introduce artifacts in the data, loss of information and generation of spurious data.

These limitations can be overcome using the LSFT, a statistical technique first introduced in astrophysics by Lomb (1976), who proposed the use of least-squares fit sinusoidal curves to the data, as a way to find periodicities in unevenly spaced data. Scargle (1982) extended Lomb's work by defining the Lomb-Scargle (LS) periodogram and by deriving the null distribution for it. Horne and Baliunas (1986) found that standardising the LS periodogram resulted in well defined statistical properties. Press and Rybicki (1989) proposed a novel fast and efficient algorithm (for its implementation see Press et al., 1992).

Consider the finite time series x_t with length T and N observations. In the case of periodic sampling, the temporal distance between two realisations is constant

$$t_j - t_{j-1} = \Delta t = \frac{1}{T}, \quad j \in \{1, 2, 3, ..., N\} \tag{1}$$

and the time series can be Fourier analysed as a sum of trigonometric functions

$$x_t = \sum_{k=-N/2}^{N/2-1} (a_k \cos(\omega_k t) + b_k \sin(\omega_k t)) = \sum_{k=-N/2}^{N/2-1} c_k e^{i\omega_k t} \tag{2}$$

where the angular frequency $\omega_k = 2\pi k/N$ and the frequency $f_k = \omega_k/2\pi$. Summing up the squared modulus of the Fourier coefficients over a specific frequency range returns an estimate of the power spectral density of the process x_t, i.e.

$$SDF_{FFT}(\omega_k) = \frac{1}{N\sigma_x^2} \left| \sum_{t=1}^{N} x_t e^{-i\omega_k t} \right|^2 \tag{3}$$

where

$$\sigma_x^2 = \frac{1}{N-1} \sum_{j=1}^{N} (x_j - \bar{x})^2$$

and the coefficients $c_k = N^{-1} \sum_{t=1}^{N} x_t e^{-i\omega_k t}$ can be computed using the well-known FFT (see, for example, Bloomfield, 2000 and Warner, 1998). However, as discussed in the previous section, UHF transaction data are observed in real-time and arrive in irregular time intervals (Δt in (1) is now stochastic), so the simple FFT can not be directly applied. The main issue here is to find an algorithm which is able to quickly compute (3).

As discussed in the previous section, this issue has been addressed in many different areas of science (for an overview, see Ware, 1998) under the robust framework of the LSFT. The general scheme is to transform the data on the non-equispaced grid into the frequency domain in order to obtain an unbiased estimation of the SDF. The corresponding Lomb-Scargle normalised SDF is evaluated at $k \in \{1, 2, 3..., M\}$ points with M chosen as outlined in Press et al. (1992). The normalised SDF is given by (for its implementation see Press et al., 1992) by

$$SDF_{LS}(\omega_k) \equiv \frac{1}{2\sigma_x^2} \left\{ \frac{\left[\sum_{j=1}^{N} (x_j - \bar{x}) \cos \omega_k (t_j - \tau)\right]^2}{\sum_{j=1}^{N} \cos^2 \omega_k (t_j - \tau)} + \frac{\left[\sum_{j=1}^{N} (x_j - \bar{x}) \sin \omega_k (t_j - \tau)\right]^2}{\sum_{j=1}^{N} \sin^2 \omega_k (t_j - \tau)} \right\} \tag{4}$$

where

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j$$

and with

$$\tau\left(\omega_k\right) = \frac{1}{2\omega_k}\arctan\left(\frac{\sum_{j=1}^{N}\sin\left(2\omega_k t_j\right)}{\sum_{j=1}^{N}\cos\left(2\omega_k t_j\right)}\right) \tag{5}$$

and $f_k = \omega_k/2\pi \in [0, 0.5]$ as the frequency. In addition, Scargle (1982, Appendix C) proved that this particular choice of the offset τ makes (4) the same equation one would obtain by linear least-squares fitting sine waves to the data (see also Van Dongen et al., 1999).

3 Simulation Results

The objective of this work is to highlight the advantages of the LSFT in obtaining the SDF for non-periodic UHF data and to demonstrate that the use of the FFT on resampled evenly spaced data leads to a SDF which includes spurious data generated by the resampling. This is achieved by employing the FFT and LSFT to analyse artificial processes with different characteristics, within a univariate framework. The simulation study is divided into the following steps, starting with the generation of a standard stationary process $\{X_t\} \sim AR(1)$,

$$X_t = \alpha X_{t-1} + \epsilon_t, \tag{6}$$

with $|\alpha| < 1$, $\{\epsilon_t\} \sim WN(0, \sigma^2)$, as the benchmark process and the computation of its theoretical normalised SDF (for the derivation of the close form solution, see e.g. Chatfield (2003) and Fan and Yao (2005)):

$$SDF(\omega) = \frac{1-\alpha^2}{\pi(1-2\alpha\cos\omega+\alpha^2)} \tag{7}$$

This benchmark process is then transformed via two resampling methods to simulate irregularly spaced data for which the LSFT is applied to calculate the SDF. These irregularly spaced data are then re-sampled to make them equally spaced and the SDF calculated using the FFT. The empirical analysis is outlined in the following steps:

(1) An evenly spaced AR(1) process is generated with parameter $\alpha = 0.9$, $N = 10,000$ data points and unitary time-span. A comparison between the SDFs of an equally-spaced AR(1) process, computed using respectively the simple FFT and the LSFT is carried out to check that both methods yield the same results.

(2) (a) By irregular resampling through N/2 random permutations the process (1), an irregular AR series is generated, with N/2 data points and variable time-span (AR-irreg). This is the first simulated UHF process. The SDF is then calculated using the LSFT.

(b) By regular resampling process (2a), an evenly-spaced process is generated, with N/5 data points and time-span equal to 5 (AR-res). The SDF is then calculated using the FFT.

(3) (a) As reported in many empirical studies in the literature, all order book data exhibit the characteristic duration clusters when computing the time interval between two consecutive economic events. Hence, in order to provide a more realistic representation of real financial market data and in particular to mimic the dynamics of a typical UHF order book data set with this stylized fact, we apply a linear ACD(1,1) model with exponentially distributed durations (Engle and Russell, 1998), yielding an irregularly-spaced process (AR-irreg-Clust) with variable time intervals between the observations of the original $AR(1)$ process. The SDF is then calculated using the LSFT.

(b) By regular resampling the clustered series of point (3a), an equally spaced series is obtained, with variable length and time-span equal to 5 (AR-res-Clust). The SDF is calculated using the FFT.

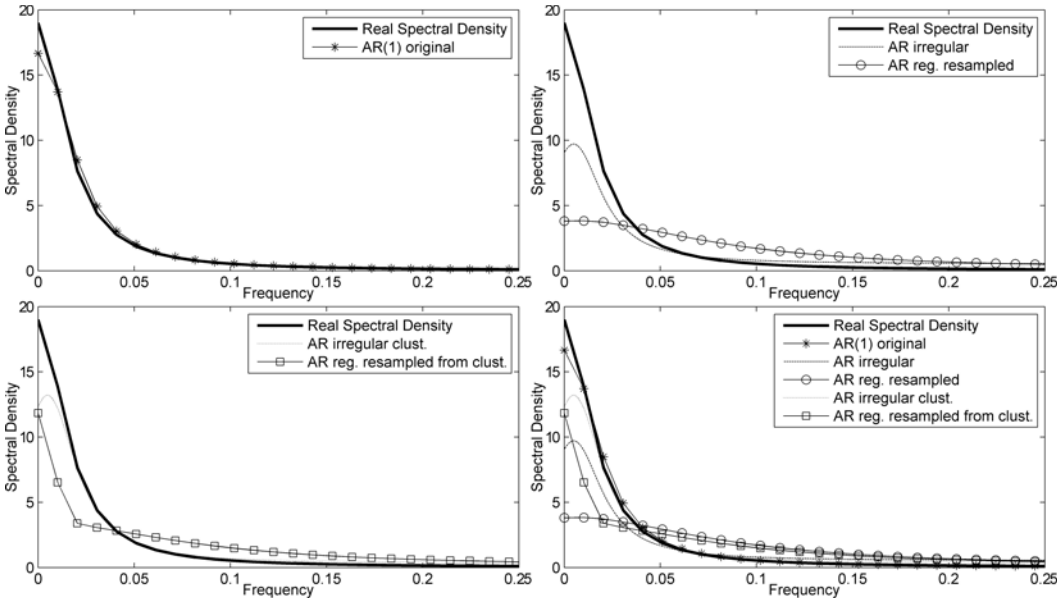
(4) Finally, a Monte Carlo simulation with $n = 10,000$ iterations is carried out repeating steps $1-3$, each time comparing the theoretical SDF (7) of the AR(1) process and the empirical SDF of the simulated processes (original AR(1), irregularly-sampled, regularly-resampled, cluster-sampled and regularly-resampled from the clustered process). Additionally, summary statistics such as the first four moments of the paired differences ("errors") between the real and the empirical spectral densities are computed.

Table 1 shows the results of the Monte Carlo simulation for a simulated series of 10,000 data points and 10,000 iterations, considering the four resampled processes (2a, 2b, 3a, 3b). An interesting point to note about the summary statistics is that while the variance of of the differences for the inhomogeneous

Table 1
Four moments of the differences ("errors") between the real spectral density and the spectral densities of the simulated processes

Data	AR1	AR-irreg	AR-res	AR-irreg-Clust	AR-res-Clust
Mean	0.0039	0.0131	0.1348	0.0509	0.0193
Variance	0.1828	2.8062	7.5068	1.0993	4.3213
Skewness	3.4664	4.4863	4.4440	5.8945	3.2496
Kurtosis	29.2934	27.5141	22.8056	38.4357	19.3971

Figure 1. A comparison between the theoretical SDF and empirical spectral densities of the simulated processes



(non-periodic) processes (2a) and (3a) is lower than that of the homogeneous (periodic) ones, the kurtosis is significantly higher.

Finally, Figure 1 depicts the results of the Monte Carlo simulation. In order to generate the graphs, we first compute for each process an average curve (of all 10,000 estimated spectral densities) and then use a filter to smooth the periodograms (Bloomfield, 2000).

Figure 1 confirms the intuition about the limitations of regular resampling introducing spurious data. As discernible in the upper left panel, the spectral densities of the autoregressive processes exhibit a declining slope, which implies a higher contribution to the variance of the process coming from low frequencies. This would translate in a significant disturbance of the processes in the long term and a lower variability in the short run. Similar results have been obtained in the literature (see e.g. Easley and O'Hara, 1992) for the traditional ACD model, in which clusters with long durations were proved to be more likely than those with short ones, as a result of asymmetric information among market participants (O'Hara, 1998).

As expected, the spectral densities of the processes (2a) and (2b) clearly demonstrate the two main problems arising from regular resampling (see upper right panel in Figure 1). At low frequencies ($f_k \leq 0.025$) the SDF of the irregular process is above the SDF of the regularly resampled one, which means that by resampling the data, we lose information that is, in fact, available. On the contrary, at higher frequencies ($0.025 \leq f_k \leq 0.225$), the reverse is true, clearly demonstrating that relying on resampled data, leads to the use of spurious information generated as a result of the resampling process (i.e. information which is actually not available). As visible in the lower left panel in Figure 1, similar considerations are also true for (3a) and (3b), but the differences have become smaller due to the sampling from clustered obeservations. For a better comparison, the lower right panel contains all spectral densities.

4 Summary and concluding remarks

UHF data are observed in real-time and therefore are characterized by the irregularity of time intervals between two consecutive events. The analysis of UHF data is now a rapidly growing area within financial research, with the prospect of providing new insights into the microstructure of financial markets. This work proposes a new approach to the study of UHF data based on the LSFT. It provides, for the first time, a rigorous theoretical framework (borrowed from other areas of science) to investigate UHF data in the frequency domain (calculation of the SDF), that incorporates, in a natural way, the non-periodic property of UHF data without the need to first transform the data to a periodic array. Any such transformation via the regular resampling of unevenly-spaced data or interpolating them to an evenly spaced grid in order to calculate the SDF with the simple FFT, is demonstrated to introduce artifacts in the data in the frequency domain (and hence time domain), leading to loss of information and the use of spurious information. The LSFT has the additional advantages of greatly reducing the computational effort required when analysing large order book data sets, avoiding complex model specifications or obligatory deseasonalisation, and providing a robust consistent framework to analyse multivariate data.

The work presented here by necessity employed simulated UHF data generated from a benchmark AR(1) process to highlight the advantages of the LSFT. The next step is to apply the theoretical framework to real order book data. In addition, the framework can be generalised to a multivariate scheme (see e.g. Schulz and Stattegger, 1997), which holds out the prospect of looking at dependencies between various variables (e.g. price, volume, etc.) that can not be easily accommodated by current econometric models in the literature. In contrast, as it is well-known in the signal processing literature, comovements of multivariate time series can be easily detected in the frequency domain by computing the cross-spectra, the gain function and the phase function. These research areas are currently under investigation.

References

Bauwens, L., Giot, P., Grammig, J., Veredas, D., 2004. A comparison of financial duration models via density forecasts. International Journal of Forecasting 20, 589–609.

Bauwens, L., Hautsch, N., 2006a. Modelling financial high frequency data using point processes.

Bauwens, L., Hautsch, N., 2006b. Stochastic conditional intensity processes. Journal of Financial Econometrics 4, 450–493.

Bloomfield, P., 2000. Fourier Analysis of Time Series: An Introduction, 2nd Edition. Wiley-Interscience.

Castiglioni, P., Rienzo, M. D., 1996. On the evaluation of heart rate spectra: the lomb periodogram. In: Computers in Cardiology 1996. pp. 505–508.

Chang, K., Monahan, K., Griffin, M., Lake, D., Moorman, J., 2001. Comparison and clinical application of frequency domain methods in analysis of neonatal heart rate time series. Annals of Biomedical Engineering 29, 764–774.

Chatfield, C., 2003. The Analysis of Time Series: An Introduction, Sixth Edition, 6th Edition. Chapman & Hall/CRC.

Dacorogna, M. M., Gençay, R., Müller, U. A., Olsen, R. B., Pictet, O. V., May 2001. An Introduction to High-Frequency Finance, 1st Edition. Academic Press.

Easley, D., O'Hara, M., 1992. Time and the process of security price adjustment. The Journal of Finance 47, 577–605.

Engle, R. F., Lunde, A., 2003. Trades and quotes: A bivariate point process. Journal of Financial Econometrics 1, 159–188.

Engle, R. F., Russell, J. R., 1998. Autoregressive conditional duration: A new model for irregularly spaced transaction data. Econometrica 66, 1127–1162.

Fan, J., Yao, Q., 2005. Nonlinear Time Series: Nonparametric and Parametric Methods. Springer.

Fernandes, M., Grammig, J., 2006. A family of autoregressive conditional duration models. Journal of Econometrics 130, 1–23.

Glynn, E. F., Chen, J., Mushegian, A. R., 2006. Detecting periodic patterns in unevenly spaced gene expression time series using lomb-scargle periodograms. Bioinformatics 22, 310–316.

Grammig, J., Heinen, A., Rengifo, E. W., 2004. Trading activity and liquidity supply in a pure limit order book market. an empirical analysis using a multivariate count data model. Working paper series, Center for Operations Research and Econometrics (CORE) - Université catholique de Louvain, cORE Discussion Paper No. 2004/58.

Granger, C. W. J., Engle, R., 1984. Applications of spectral analysis in econometrics. Vol. 3 of Handbook of Statistics. Elsevier Science Publishers, Ch. 5, pp. 93–109.

Hall, A., Hautsch, N., 2006. Order aggressiveness and order book dynamics. Empirical Economics 30, 973–1005.

Horne, J. H., Baliunas, S. L., 1986. A prescription for period analysis of unevenly sampled time series. Astrophysical Journal 302, 757–763.

Iacobucci, A., 2003. Spectral analysis for economic time series.

Laguna, P., Moody, G., Mark, R., 1998. Power spectral density of unevenly sampled data by least-square analysis: performance and application to heart rate signals. Biomedical Engineering, IEEE Transactions on 45 (6), 698–715.

Lomb, N. R., 1976. Least-squares frequency analysis of unequally spaced data. Astrophysics and Space Science 39, 447–462.

Muller, R. A., Macdonald, G. J., 2000. Ice Ages and Astronomical Causes: Data, Spectral Analysis, and Mechanisms. Springer-Verlag Telos.

O'Hara, M., 1998. Market Microstructure Theory. Wiley-Blackwell.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T., 1992. Numerical Recipes in C: The Art of Scientific Computing, 2nd Edition. Cambridge University Press.

Press, W. H., Rybicki, G. B., 1989. Fast algorithm for spectral analysis of unevenly sampled data. Astrophysical Journal 338, 277–280.

Priestley, M. B., 1981. Spectral Analysis and Time Series. Volume 1: Univariate Series., 1st Edition. Academic Press.

Ruf, T., 1999. The lomb-scargle periodogram in biological rhythm research: Analysis of incomplete and unequally spaced time-series. Biological Rhythm Research 30, 178.

Scargle, J. D., 1982. Studies in astronomical time series analysis. ii - statistical aspects of spectral analysis of unevenly spaced data. Astrophysical Journal 263, 835–853.

Scargle, J. D., 1989. Studies in astronomical time series analysis. iii - fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data. Astrophysical Journal 343, 874–887.

Schulz, M., Stattegger, K., 1997. Spectrum: spectral analysis of unevenly spaced paleoclimatic time series. Computers & Geosciences 23, 929–945.

Van Dongen, H., Olofsen, E., Van Hartevelt, J., Kruyt, E., 1999. A procedure of multiple period searching in unequally spaced time-series with the lomb-scargle method. Biological Rhythm Research 30, 149–177.

Ware, A. F., 1998. Fast approximate fourier transforms for irregularly spaced data. SIAM Review 40, 838–856.

Warner, R. M., 1998. Spectral Analysis of Time-Series Data, 1st Edition. The Guilford Press.